家庭内暴力の解析のための位相マップ

Jonas Poelmans[1] , Marc M. Van Hulle[5] , Paul Elzinga[2], Stijn Viaene[1,3], Guido Dedene[1,4]

[1]K.U.Leuven, Faculty of Business and Economics, Naamsestraat 69,
3000 Leuven, Belgium
[2]Police Organisation Amsterdam-Amstelland, James Wattstraat 84,
1000 CG Amsterdam, The Netherlands
[3]Vlerick Leuven Gent Management School, Vlamingenstraat 83,
3000 Leuven, Belgium
[4]Universiteit van Amsterdam Business School, Roetersstraat 11
1018 WB Amsterdam, The Netherlands
[5]K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg O&N2, Bus 1021, Herestraat 49
3000 Leuven, Belgium
{Jonas.Poelmans, Stijn.Viaene, Guido.Dedene}@econ.kuleuven.be
Paul.Elzinga@amsterdam.politie.nl
marc@neuro.kuleuven.be

アブストラクト．位相マップは，データベースから新しい知見を発見するための探索的な興味を引くツールである。ここしばらくの間に，新しい形式の自己組織化マップ(SOM)が文献に発表されている。この章では，新しいエマージェント SOM や新しい球面 SOM が興味あるデータセットを研究するために使われている。データセットは，アムステルダム-アムステランド(オランダ)の警察地域で 2006 年の第第 1 四半期に起こった全ての暴力的な出来事を記述している警察調書のまとめから成っている。球面位相マップは，このデータセットを解析するための強力なツールとして準備されている。さらに，球形の地形図が強力な機器を提供することは示される。さらに，エマージェント SOM の性能は，球面 SOM の性能と比較された。そして，直接データに適用して見て，普通のクラス分類器の性能よりも優れていることが判明した。

キーワード：位相マップ，家庭内暴力，データベースからの知識発見，エマージェント SOM，BLOSSOM

1 はじめに

オランダ法務省の報告によると，家庭内暴力は“犠牲者の家族の誰かによって犯された重大な暴力行為”と見なすことができる。暴力とは暴力行為の全ての形式を見ている。家族の範囲は，犠牲者の全てのパートナー，過去のパートナー，家族，親族，並びに家族一同の友人達を含んでいる。“家族一同の友人とは，犠牲者との友好的な関係を持っていて，規則的に，彼/彼女の家で犠牲者と会う人達である”[1]。

研究の結果，家庭内暴力は，我々の現代社会では，非常に過小評価された問題であることが判明した [2,3,4,5]。犯罪者に対して効果的な方針を実行することはオランダの地域アムステルダム-アムステランドの警察組織の最優先事項の 1 つである。もちろん，犯罪者に対して効果的な方針を実行するために，速やかに家庭内暴力のケースを認め，それに応じてレポートにラベルを貼って整理できることは最も重要なことである。それでも，これは，疑わしいと判明した。過去に，ファイル整理されたレポートに関連した警察データベースへの集中的な監査の結果，多くのレポートが，家庭内または非家庭内の暴力事件の場合として間違って分類される傾向があったことを立証した。結論の１つは，この問題領域の詳細な調査の必要が出てきたということである。

この論文では，特に高次元のデータの視覚化に適している位相マップ[7]を使って，この問題領域は研究されるであろう。ここでは，最近の２つのツールが考慮されている。１つは，エマージェント SOM で，もう１つは球面 SOM であり，その性能を比較した。

2　位相マップとして必要なこと

実践的な観点からは，位相マップは，データベースでの知識発見のための特に興味を引く技術である[15]。それは低い次元のもの，通常 2 次元のものへと高次元空間の非線形なマッピングを実行する。データセットを探究するために，有益なツールがユーザーに提供されている[12]。それは，クラスタを検出するために使われることができて，入力空間に存在する近傍関係を維持しているものである。それはまたユーザーにデータセットの複雑さ，データセット（例えば球形の場合）での分布，および異なったクラス間の重なり合いの量とかについての考え方を提供する。より低い次元でのデータ表現は，またクラス分類器を組み立てる時に有効である。

2.1　エマージェント SOM

エマージェント自己組織化マップ(ESOM)は，ごく最近現れた位相マップである[8]。それは，その構造の直観的な概要を産み出して，まばらで，高次元のデータセットを視覚化することに特に有用であると議論されている[10]。エマージェント SOM は以下の点で通常の SOM と異なっている。そこでは，大変大きい数のニューロン（少なくとも数千個）が使われる[9]。通常の SOM での射影の位相保存は，小さなマップを使う時にほとんど役立っていないと言われている：小さな SOM の性能は，k をマップ内のノードの数に等しくした k-ミーンズ・クラスタ分析とその結果がほとんど同一であると議論されている。ESOM の別の利点を以下に記述する。それは，最初に，特徴選択の手順を実行しないで入手可能なデータセット上で直接学習することができる[11]。ESOM マップは公開されて入手可能な *Databionics ESOM* ツールによってデータ解析のために作成されて，使用できる。このツールは、ユーザーが平面 SOM と，境界が無い（すなわちドーナツ形をしている）ESOM マップの両方を組み立てることを可能にする。

2.2　球面 SOM

球面 SOM では，ニューロンは球面に配置される。最近，いくつかの球面自己組織化マップが文献に紹介されている[6]。これらのマップは，球面でも，また，ドーナツ型（トロイダル型 SOM）でも，例えば通常の SOM や，その多くの改定版の場合のように，境界が無い。そして，このため，これらのマップは境界，周辺効果を受けるべきでは無い。マップの境界での１つのニューロンの近傍ニューロンの数がマップの中心での１つのニューロンの近傍ニューロンの数より少ないので，境界，周辺効果は，通常の平面マップで起きる現象である[14]。これは，マップの歪みを起こし，例えばマップの端近くでは，クラスタ検出をするためには，あまりにも小さな領域をもたらすかもしれない。ここで使われた球面 SOM ツールは、BLOSSOM [13]であるけれども，それはまだ，高次元のデータセットには適用されていない。

3　データセット

データセットは 2006 年の第 1 四半期からすべての暴力事件を記述している 4146 件の警察調書から成っている。その期間での全ての家庭内暴力事件はこのデータセットの 1 部分を形成している。あいにく，これらの 4146 の警察の調書の多くは，家庭内暴力と認定するのに必要である犠牲者による犯罪の報告を含んでいなかった。つまりこう言うことである。例えば、警察官が事件現場に派遣されて，その後，彼／彼女が彼／彼女が見たことについて言及したリポートを書くと言う形で事件は整理される。一方，犠牲者は警察に対して公式な記述をしてこなかった。従って，我々は犠牲者が犯罪事件を警察官に報告した，たった 2288 の調書を確保できただけであった。これらの 2288 の調書から，我々は前の事件に関係する続報を取り除いた。この精査で，結果として 1794 の調書が残った。これらのリポートから，犯罪を報告した人，容疑者，犯罪に関係している人，目撃証人，プロジェクトコード，および声明を引き出した。これらの 1794 の調書の中から，462 件が家庭内暴力に関係する調書であっ

た；そして他は関係が無かった。これらのデータは，研究で使われた 1794 の html 文書を生成するために使われた。そのようなリポートの 1 例は図 1 に示す。

事件名	事件 xxx
調書日	26-11-2007
プロジェクトコード	年長者に対する家庭内暴力(+55)
事件場所	アムステルダム　カイザーグラフト　yyy
被疑者（男）(18-45 歳)	zzz
住所	アムステルダム　カイザーグラフト　yyy
事件に巻き込まれた人（男）(18-45 歳)	隣人
住所	アムステルダム　カイザーグラフト　www
被害者（男）(>45 歳)	uuu
住所	アムステルダム　カイザーグラフト　vvv

事件の報告

昨晩、私は私のたった 1 人の息子に襲われた。私がリビング・ルームでテレビを見ていた時に，彼が突然ナイフで襲ってきた。私はフロアに倒れた。それから，彼は，私を蹴ろうとした。私は助けを求めて叫びながら，私は，裏口から脱出しようとした。私は助けを求めて隣人宅に逃げた。隣人は救急車を呼んだ。その間に，私の息子は逃走した。私の足は出血していた。．．．

図 1. 警察調書の例

私達はまた自由に事項別語彙集を持っている－用語集－それは，これらの警察調書についての頻度分析をすることによって得られたものである。非常にしばしば出てくる用語は取り出され，最初は空である事項別語彙集に追加された。これは結果として 123 の用語のセットを生じていた。

順序付け出来ないデータセットの中で，これらの用語のどれがリポートに出現するかは各警察の調書で示される。このデータセットの抜粋は表 1 に示される。

表 1. 研究に使われた順序付け出来ないデータセットの抜粋。

	蹴る	父親の暴力	突き刺し	呪う	ひっかく	虐待
レポート 1	X	X				X
レポート 2			X	X	X	
レポート 3	X	X	X	X	X	
レポート 4						X
レポート 5				X	X	

連続数で表せるデータセットの中で，各用語の関連は 0 から 1 の間で連続値によって各警察で示される。この値は，用語がリポートに出現した数に基づいて計算された。

表 2. 研究に使われた連続数で表せるデータセットの抜粋。

	蹴る	父親の暴力	突き刺し	呪う	ひっかく	虐待
レポート 1	0.371	0.781	0	0	0	0.581
レポート 2	0	0	0.02	0.83	0.496	0
レポート 3	0.238	0.421	0.387	0.628	0.921	0
レポート 4	0	0	0	0	0	0.862
レポート 5	0	0	0	0.782	0.4	0

各警察の調書にとって，いくらかの追加情報は入手可能である。刑事犯罪の容疑者が知られているかどうか，犠牲者の性，犠牲者の年齢，犯人と犠牲者が同じ住所に住んでいたかどうか，等の情報を含んでいる。

まず始めに，ドーナツ形の ESOM マップは，これら 2 つのデータ集合を使って，データ集合の分布を見るために訓練された。図 2 で表示されたマップの中で，最整合（最近接）ノードは，与えられたテストデータ集合（家庭内暴力のための赤，非家庭内の暴力のための緑）のために 2 クラスにラベル付けされている。順序付け出来ないデータ集合に基づいて ESOM マップを解析することによって，データは球状に分布していると結論するのは可能になった。マップの中央を通って垂直に走り抜けている 1 つの大きな家庭内暴力のクラスタとその左に走っている少しぼんやりと境界設定された家庭内暴力のクラスタがあることを見つけることができる。後者はマップの端へと続いていて，マップの右側に外れ値がある。従って，このデータ集合を可視化するためには球面 SOM を使うことは自然な流れである。

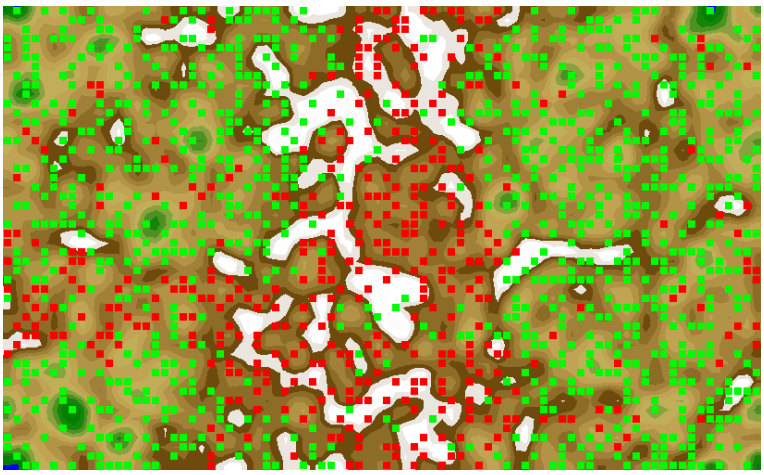

図 2. すべての特徴を持つ順序付け出来ないデータ集合で学習されたドーナツ型 ESOM マップ。

両方のツールにとっては，123 要素を超える全データ集合でマップを直接作成することは可能であった。しかし，無関係で，冗長な特徴によって引き起こされたマップの上の歪みを防ぐために，特徴選択を適用することにした。参考文献[16]に記述されているように最小冗長性-最大関連性基準(mRMR)として知られている発見的な特徴選択手段を利用した。目的は，最も関連した 50 の特徴を選ぶことであった。最適な特徴セットを得るために，SVM，ニューラルネットワーク，kNN(k=3 を持つ)，および単純ベイズ分類器の Naïve ベイズ選別器は，増大する特徴の分類性能を測定するように使われた。分類性能は図 3 中の特徴数の関数としてプロットしている。

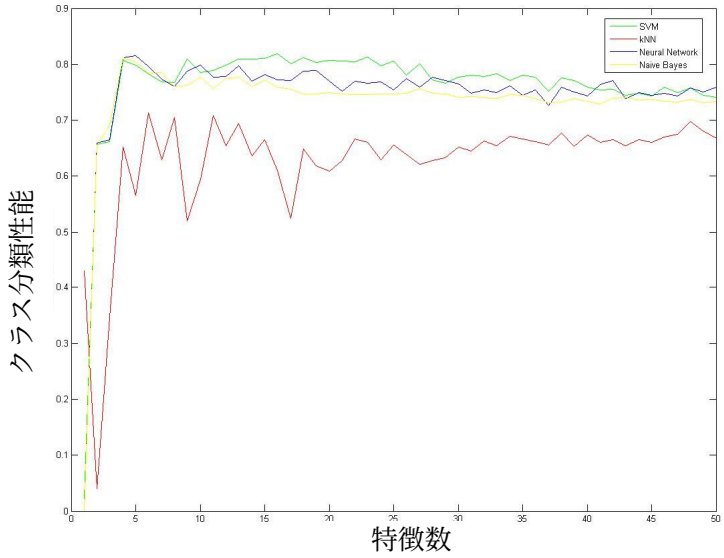

図 3. 分類性能

最もよい 18 の特徴を保持することが決定された。ESOM のために，50 行×82 列のニューロンを持つ SOM が使われた（そこでは全ニューロン数は 50×82=4100 ニューロンである)。荷重は，対応する特徴と同じ平均と標準偏差を持つガウス性をサンプリングすることによって乱数で初期化された。初期半径 24 を持つガウス型ベル形のカーネルが近傍関数として使われた。さらに，初期学習率係数は 0.5 で，この学習率係数を線形に減少させる方法が使われた。学習回数は 20 に設定された。平面位相と同様に，ニューロンのドーナツ形位相を持つ両方のマップが使われた。BLOSSOM では，642 個のニューロンから成るネットワークが使われた。荷重は乱数で初期化された。初期半径πを持つガウス型のカーネルが近傍関数として使われた。さらに，初期学習率係数は 0.9 で，この学習率係数を線形に減少させる方法が使われた。そして学習回数は 50 に設定された。

順序付け出来ないデータ集合で学習された BLOSSOM マップは図 4 に表示される。連続数で表せるデータセットでの BLOSSOM マップは図 5 に示される。順序付け出来ないデータ集合で学習されたドーナツ型 ESOM マップは図 6 に示される。順序付け出来ないデータ集合で学習された平面型 ESOM マップは図 7 に示される。

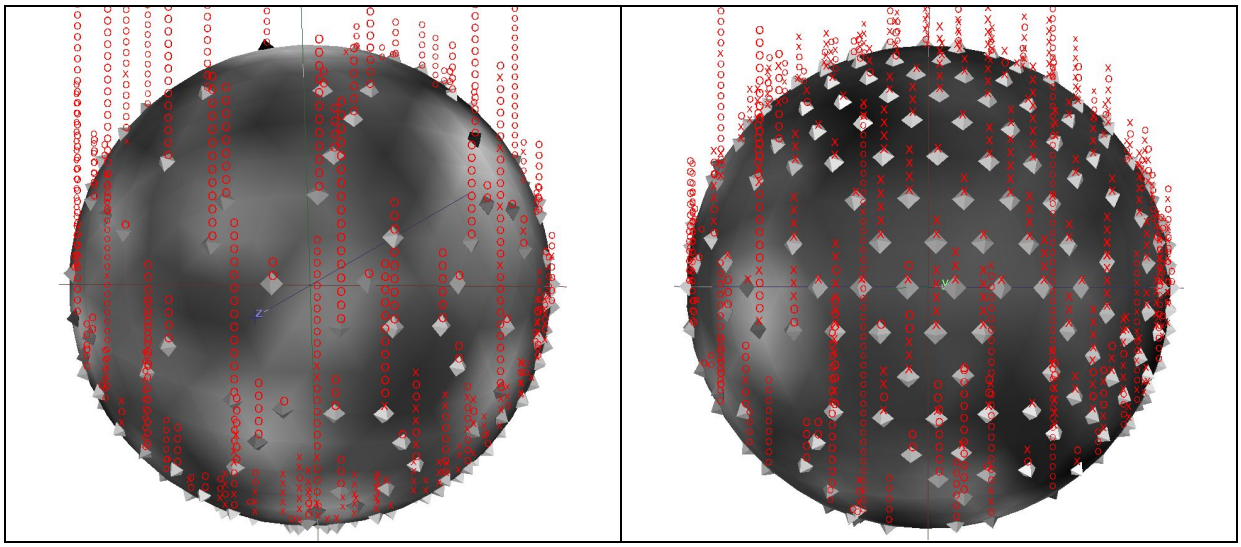

図 4. 18 次元の順序付け出来ないデータ集合で学習された 2 つの BLOSSOM マップ。球面上のグレースケールは局所密度を示す（白= 高密度）。小さな四面体は 2 タイプのラベルの最近接近傍のニューロンを示す；“ x ”は家庭内暴力，“ o ”は非家庭内暴力を示す。

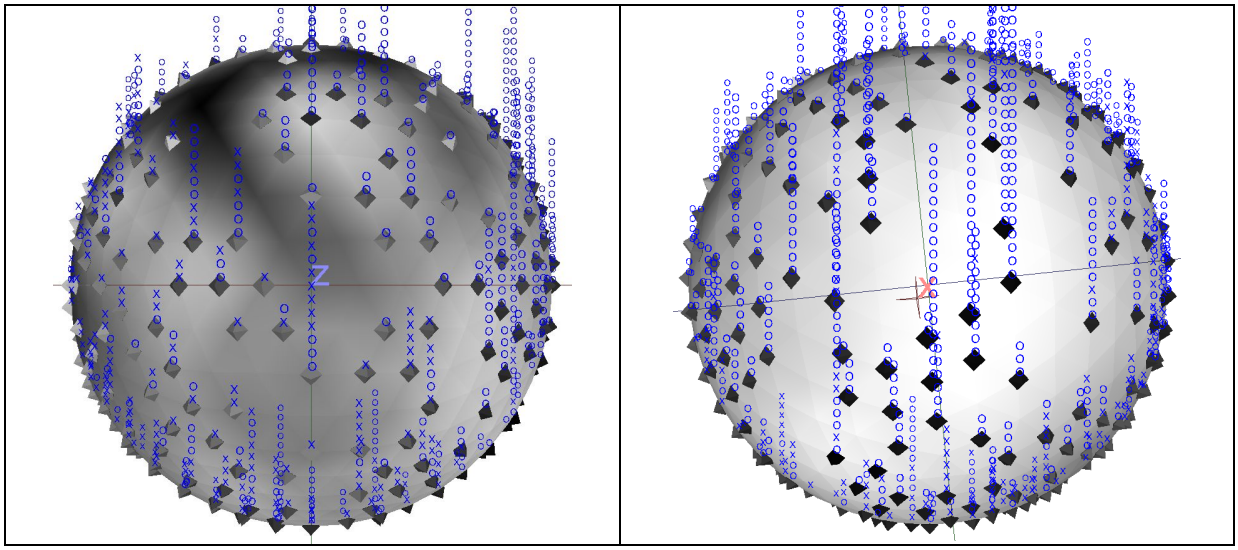

図 5. 18 次元の連続数で表せるデータセットで学習された 2 つの BLOSSOM マップ。

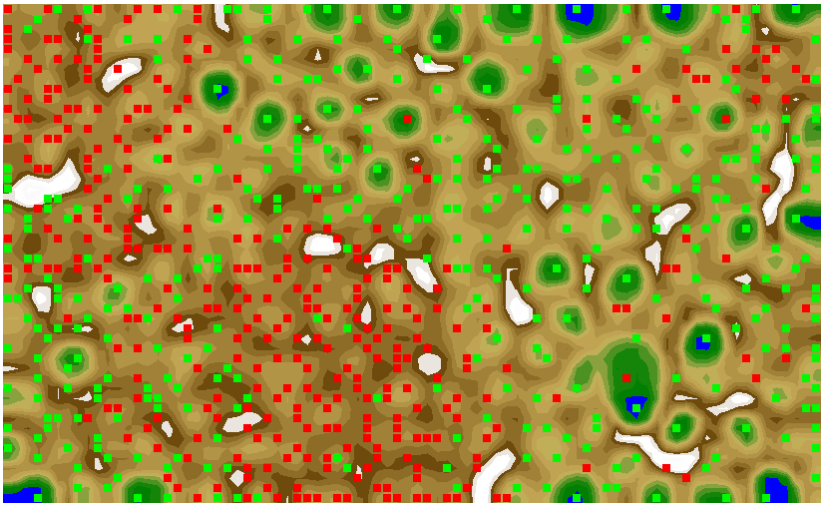

図 6. 18 次元の順序付け出来ないデータ集合で学習された平面 ESOM マップ。

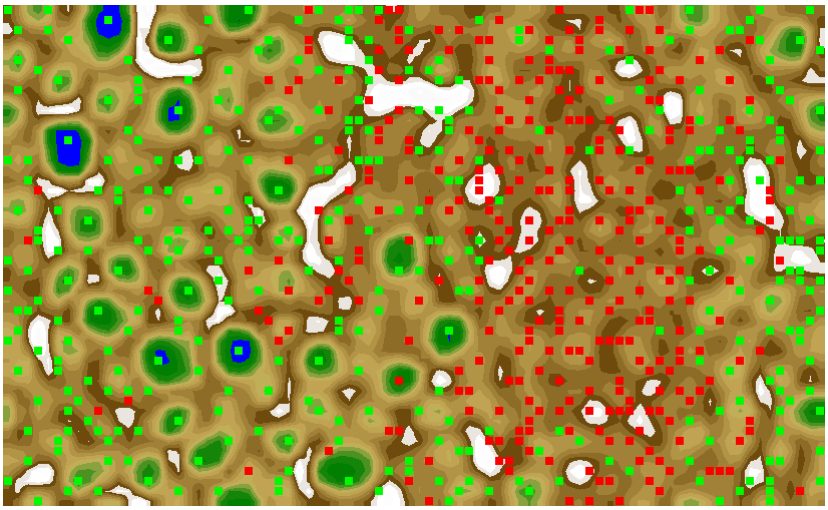

図 7. 18 次元の順序付け出来ないデータ集合で学習されたドーナツ型 ESOM マップ。

最後に，ESOM と BLOSSOM マップのために kNN クラス分類器が構築された。BLOSSOM では，k は 1 に設定された。BLOSSOM マップでの分類誤り誤差を得るために，各荷重ベクトルへの各入力ベクトルのユークリッド距離が測定された。（マップのノードに対応している）各荷重ベクトルにとって，家庭内及び非家庭内暴力のケースのいくつが最整合の荷重ベクトルを持つかが計算された。ノードが家庭内暴力ケースを支配的に占めている場合には，このノードは家庭内暴力ノードとラベル付けし，このノードに最整合する非家庭内暴力のケースは，間違った分類であると考えた。ESOM マップが BLOSSOM マップの約 2 倍の最整合ニューロンを含んでいたので(680 対 316)、k は ESOM マップでは 2 に設定された。最整合ニューロンとは，このノードの荷重ベクトルへのユークリッド距離が最小の少なくとも 1 つの入力ベクトルが存在しているニューロンのことである。

4　結果

ESOM ツールは，データ集合の構造の概要を迅速に掴むために全ての特徴を含んでいるデータ集合にまず適用された。これは，ユーザーが容易にデータ集合の球面特性を検出し，従って BLOSSOM を興味深い候補ツールにすることを可能にした。ESOM マップについての問題点は，マップの密度輪郭がラベル付けされたデータベクトルの一様分布と整合しないことである。さらに，マップ中で家庭内暴力と非家庭内暴力の領域を分離する山が見えない。従って，‘ watershed（分岐）’技法[9]はクラスの

正しい識別をもたらさないであろう。この問題は，特徴の数を減らすことによっても解決されなかった。それにもかかわらず，ずっと多くの密度変化が図 7 中で観察することができる。BLOSSOM ツールで遭遇する主要な問題点は，それがただ制限された次元数によってデータ集合を準備しなければならないことであった。BLOSSOM はこの大きさのデータ集合に初めて適用された。全てのラベルがマップ上で見えるわけではないことは問題であった。マップの上部のラベルの多くは，ツールに使用された小さな窓サイズのために見ることが不可能であった。

図 4 に示す BLOSSOM マップを調べることによって，順序付け出来ないデータ集合では，家庭内暴力の場合に，はっきりと境界設定されたクラスタが全然入手可能でないと断定することができる。これはまた，2006 年の第 1 四半期での家庭内暴力の場合だけを含んでいるデータ集合で学習された球面マップの場合であった。後者は，小さい密度変化だけを含んでいる。しかし，非家庭内暴力の場合には，いくつかのはっきりと境界設定されたクラスタが図 4 のマップ中で見つけることができる。

これらのクラスタは多分，他の型の事件とはっきりと区別することができる違う型のものと一致している。強盗事件では，例えば容疑者は，一般には知られず，容疑者の記述も提供されないし，家の中の 1 つ以上の位置も言及されていない。これらの典型的な特徴は，強盗事件を認めることを容易にする。

図 5 中に示されたマップから，これがまた連続数で表せるデータ集合の場合であると断定することができる。しかし、後者がずっと少ない密度変化を含んでいることが目立っている。

別の興味深い結果は，マップが家庭内暴力と非家庭内暴力の間でよい分別を提供していることである。最整合のノードの多くは，支配的に家庭内暴力，または非家庭内暴力の場合を含んでいる。これは，それらの間にほんの少量のオーバーラップがあることを示す。観察されたオーバーラップは多分，特徴の集合が，2 クラスを区別するために十分に改良されていないことを示す。しかし，いくつかのケースは警察官によって間違って分類されていたかもしれないことは考慮されるべきである。後者は家庭内暴力の定義のあいまいさによるかもしれない。

平面 ESOM マップとドーナツ形 ESOM マップを比較すると，ドーナツ形マップの方がデータ集合の視覚化にはずっとよい結果を提供すると結論できる。平面マップには，境界効果の影響で，マップの望ましくない歪みが結果として生じていることがはっきりと分かる。観察されたクラスタのほとんどは，マップの境界に置かれる。このことは，その領域をより小さくする。そして，家庭内暴力の場合の大きなグループは，非家庭内暴力の場合との間に，よりぼんやりとした境界を示される。

最後に，ESOM と BLOSSOM マップに基づいた最近接分類器の結果は，表 3 と 4 に示される。

表 3. 順序付け出来ないデータ集合での分類性能。

	全体の精度	偽陽性率	偽陰性率
BLOSSOM 1NN	90.4%	2.9%	29%
ドーナツ型 ESOM 2NN	90.8%	8%	12.6%
平面 ESOM 2NN	91.4%	6.8%	13.9%
ドーナツ型 ESOM 1NN	95%	4%	7.6%
平面 ESOM 1NN	95.1%	4%	7.4%

興味深い結果は，通常の kNN 分類器(65%)と球面 BLOSSOM マップに基づいた kNN 分類器(90.4%)の性能における著しい違いである。これはデータ分布のモデルである位相マップが原因である：そして，内挿の技術を使いながら，データの特徴の近似を形成し，局所密度のモデル化を導いて，データ内のクラスタ分類に，より多くのニューラル計算を行う。

表 4. 連続数で表せるデータ集合での分類性能。

	全体の精度	偽陽性率	偽陰性率
BLOSSOM 1NN	87%	3.7%	40%
ドーナツ型 ESOM 2NN	88.7%	8.6%	20.6%
平面 ESOM 2NN	88.9%	8.8%	18%
ドーナツ型 ESOM 1NN	94.4%	0.3%	21%
平面 ESOM 1NN	94.7%	0.3%	20%

表 3 と表 4 から，BLOSSOM マップに基づいた 1NN 選別器の全体の精度と ESOM マップに基づいた 2NN 選別器の全体の精度がほとんど等しいと断定することができる。しかし，偽陽性率と偽陰性率には明白な差を観察することができる。BLOSSOM マップでの偽陽性率（すなわち，家庭内暴力とクラス分けされた非家庭内暴力の数をデータ集合に含まれた非家庭内暴力の数で割ったもの）は，ESOM マップでの偽陽性率よりも 2 倍良い。その反対に，偽陰性率（すなわち，NN クラス分類器によって家庭内暴力と分類されなかった家庭内暴力の数を，データ集合の中にある家庭内暴力の数で割ったもの）に対して真である。驚いたことに，平面とドーナツ型 ESOM マップの間には分類性能における違いはほとんどない。前者（平面マップ）では多くの望ましくない歪みを含んでいるけれども，この歪みはより低い分類精度の結果をもたらさない。別の興味深い結果がある。それは，連続数で表せるデータ集合での全体の分類精度が少しより悪いだけなのに，両方のデータ集合の偽陰性率間には非常に大きな違いがある。偽陰性率の方が重大なので，ESOM マップは BLOSSOM に比べて我々の場合にはより適していると言える。最後に以下の注目すべきことが言える。SVM などのより複雑なクラス分類器は ESOM または BLOSSOM より性能は良くなかった。ここで使われた現システムは多層パーセプトロンであるが，これは（つまり，構造がブラックボックスであるので）問題の中味が見えてこない。そして，その性能も大体たったの 80%位である。

5　結論とこれからの仕事

この章では，興味深い警察のデータセットを研究するために 2 つの最近の SOM ツールの有用性が提示された。ESOM ツールを適用することによって，データ集合の分布が球面であることを見つけることは可能であった。結果として，球面 SOM ツール BLOSSOM は，データに自然に適用できると思われる。この球面 SOM 技術を使って，データ探索の目的での興味深い結果が見つかった。最後に，ESOM と BLOSSOM マップ間の比較は，最近接近傍のクラス分類器の方法で実行された。

しかし，ESOM と BLOSSOM ツールを，完全な 123 次元の全要素を使った本格的なベンチマークデータで調べることは，この章の範囲を越えており，これは，将来にわたっての研究のトピックであり，注目すべきことである。

謝辞

著者等は，Laboratorium voor Neuro- en Psychofysiologie KULeuven の Gert Van Dijck 氏には，特徴選択についての彼のアドバイスに対して，また，アムステルダム Amstelland 警察組織に対しては，我々にデータを供給して頂いたことに対して感謝している。著者の JP は，警察組織アムステルダム Amstelland によって，著者の SV に与えられた研究ポジションによってサポートされている。

参考文献

[1] Keus, R., Kruijff, M.S. (2000) Huiselijk geweld, draaiboek voor de aanpak. Directie Preventie, Jeugd en Sanctiebeleid van de Nederlandse justitie.
[2] Watts, C., Timmerman, C.: Violence against women: global scope and magnitude. The Lancet 359 (9313): pp.1232-1237. RMID 1155557
[3] Waits, K. (1984-1985). The criminal Justice System’s response to Battering: Understanding the problem, forging the solutions. Washington Law Review 60: pp. 267-330
[4] Catriona Minleer-Black (1999) Domestic violence: Findings from a new British Crime Survey self-completion questionnaire. London: Home Office Research Study.
[5] Vincent, J.P., Jouriles, E.N. (2000) Domestic violence. Guidelines for research-informed practice. Jessica Kingsley Publishers London and Philadelphia
[6] Ritter, H. (1999) Non-Euclidean Self-Organizing Maps, pages 97–109. Elsevier, Amsterdam.
[7] Kohonen, T. (1982), “Self-Organized formation of topologically correct feature maps”, Biological Cybernetics, Vol. 43, pp 59-69.
[8] Ultsch, A., Moerchen, F. (2005) ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany, No. 46

[9] Ultsch, A., Hermann, L. (2005) Architecture of emergent self-organizing maps to reduce projection errors. In Proc. ESANN 2005, PP1-6
[10] Ultsch, A. (2004) Density Estimation and Visualization for Data containing Clusters of unknown Structure. In proc. GfKI 2004 Dortmund, pp 232-239
[11] Ultsch, A. (2003) Maps for visualization of high-dimensional Data Spaces. In proc. WSOM'03, Kyushu, Japan, pp. 225-230
[12] Ultsch, A., Siemon, H.P. (1990) Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis. Proc. Intl. Neural Networks Conf., pp305-308
[13] Tokutaka, H., BLOSSOM Software Tool, http: //www.somj.com
[14] Nakatsuka, D., Oyabu, M. (2003) Application of Spherical SOM in Clustering. Proc. Workshop on Self-Organizing Maps (WSOM '03), pp 203-207
[15] Van Hulle, M. (2000) Faithful Representations and Topographic Maps from distortion based to information based Self-Organization. Wiley: New York
[16] Peng, H., Long, F., Ding, C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependancy, Max-Relevance, and Min-Redundancy. IEEE Tansactions on pattern analysis and machine intelligence, Vol. 27, no. 8.